Three public standards development projects are delivering the resources needed to support truly integrative, robust data management. But what are the potential benefits for the pharmaceutical industry, and at what cost?

Standards for reporting bioscience data: a forward look

Chris F. Taylor

European Bioinformatics Institute, Hinxton, Cambridgeshire CB10 1SD, UK

Groups representing a number of domains in the life sciences have been developing specifications and resources for the description and transmission of data, including those produced by (high-throughput) omics technologies. Although these developments are individually valuable, there is now a need for coordination to avoid the problem of a multiplicity of competing candidate standards. Three ongoing collaborative projects (FuGE, OBI and MIBBI) offer the promise of support for truly integrated, cross-domain informatics solutions. This article briefly summarizes the *status quo* with respect to biological and biomedical data standards, and offers an assessment of coming developments.

As laboratory workflows become ever more diverse and prolific and multi-domain (i.e. systems biology) analyses become more commonplace, the efficient capture and effective analysis of data increasingly require solutions ('reporting structures') that facilitate the integration of a focused set of information from diverse domains; for example, enabling the consistent description of the key data from transcriptomic, proteomic and metabolomic analyses of the same source material, thereby allowing those data to be combined and managed as a unit without undue effort.

For public sector science, decisions regarding keeping information from experiments and the mechanisms by which that should be achieved are usually *ad hoc*. For the private sector, ten years on from the advent of 21 CFR Part 11, keeping all available information is *de rigueur*, but that requirement can be met through the use of any number of infrastructures (often comprising a slew of incompatible systems). The esoteric, heterogeneous, high-maintenance solutions that result (in the private and public sectors alike) often prevent data from rendering their full value, largely because of the expenditure of time and effort required to retrieve and reanalyze them. For industry, this constitutes a loss of 'business intelligence' (BI); that is it hinders a company's ability to establish what it already knows about something.

The opinion of public sector community opinion [1,2] increasingly favors that a prescribed subset of the available metadata ('data about the data') pertaining to an experiment be associated with the statement of its results (i.e. data and conclusions), making explicit both the biological and the methodological context; such prescriptions are known as minimum information (MI) specifications. Furthermore, (meta)data should be captured in standard non-proprietary formats, and unambiguous, standard ('controlled vocabulary') terms should be used wherever it is feasible to do so.

For the private sector, the use of standard formats and controlled vocabularies offers the same benefits as it does for the public sector; data are more easily retrieved, analyzed, compared and integrated, largely because access to them is not contingent upon the use of an arsenal of (potentially obsolescent) software tools. MI specifications also have relevance for both sectors, but the drivers differ significantly. For public science (where there is no equivalent of things like 21 CFR Part 11), the aim is simply to raise the bar with respect to the reporting of research (in the literature and through public repositories). The benefits of this are intuitive and have been demonstrated [7,8]. For the private sector, the issue is one of efficiency: Capturing a *reduced* set of metadata in a rigorous way (in addition to archiving the full set) facilitates efficient retrieval, reanalysis and integration of data. This is a 'can't see the wood for the trees' argument, the validity of which is demonstrated by the abundance of sophisticated software designed to distill all-encompassing data sets into useful summary reports.

The three basic components of a modern reporting structure (MI specifications, data formats and controlled vocabularies) are

E-mail address: chris.taylor@ebi.ac.uk.

1359-6446/06/$ - see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2007.05.006

further described in the next section; examples of each are also given in Table 1, which lists a number of publicly available resources relevant to the reporting of life sciences experimentation (note that not all are 'omics-focused' *per se*).

Good things come in threes

As stated above, regulatory requirements mandate that all the information generated in the course of an investigation be retained. By contrast, MI specifications require that only specific subsets [3,4] of the total available information pertaining to an experiment be recorded; for example, on the provenance of the study material or the manner of use of a particular analytical platform. It is the strong opinion of the author that to have general utility, such specifications should not stray into prescribing either how work should be performed (something better left to documents akin to [5,6]) or how information should be transmitted (although supplementary guidance could be provided on such matters). It should be noted that although the (automated) creation of reports complying with MI specifications is straightforward, *validating* compliance in an automated fashion is not. Currently, such checking can only be done effectively by hand, as ArrayExpress [9] and the Gene Expression Omnibus [4] have now pledged to do with respect to Minimum Information About a Microarray Experiment (MIAME) [10] compliance.

Controlled vocabularies (CVs) contain sets of words or phrases representing particular entities, processes or abstract concepts. CVs structured as hierarchies of like terms are known as ontologies. Individual terms are usually associated with a unique (within the particular CV) identifier and an unambiguous definition, ideally with examples of usage. Terms can also carry lists of (exact) synonyms, reducing the scope for misinterpretation and enabling more effective repository searches. A standing challenge for software developers is to produce tools that enable the straightforward use of CV terms by scientists doing primary data entry (especially as free text), which in addition to the above benefits helps to reduce the number of typographical errors that occur during data entry: an important problem, as database searches are usually performed with *correctly spelled* terms.

Data formats are the containers with which to transmit information from data entry or instrument control software to analysis/visualization software or repositories (which may in reality be little more than indexed archives of such files). Data formats serve many purposes, but for a format to have general utility, it should support both rich descriptions replete with metadata and skeletal descrip-

TABLE 1

Some of the publicly available resources relevant to the reporting of life sciences experimentation

Product name	Scope	Developers	Link
(i) Minimum information specifications			
SEND	Animal toxicology	CDISC	http://www.cdisc.org/models/send/v2.3/
CIMR[a]	Metabolomics	Metabolomics Standards Initiative	http://msi-workgroups.sourceforge.net/
CONSORT[b]	Clinical trials	CONSORT Group	http://www.consort-statement.org/
MCP guidelines[c]	Proteomics	*Ad hoc* collaboration	http://www.mcponline.org/misc/ParisReport_Final.shtml
MIACA	Cell-based assays	Cell-Based Assay Standards Consortium	http://miaca.sourceforge.net/
MIAME/Env	Environmental genomics	RSBI Environmental Genomics WG	http://envgen.nox.ac.uk/miame/miame_env.html
MIAME/Nutr	Nutrigenomics	RSBI Nutrigenomics WG	http://www.mged.org/Workgroups/rsbi/rsbi.html
MIAME/plant[d]	Phytology	*Ad hoc* collaboration	http://dx.doi.org/10.1186/1746-4811-2-1
MIAME/tox	Toxicogenomics	RSBI Toxicogenomics WG	http://www.mged.org/Workgroups/rsbi/rsbi.html
MIAPA[d]	Phylogenetic analyses	*Ad hoc* collaboration	http://dx.doi.org/10.1002/pmic.200500856
MIAPE[a]	Proteomics	HUPO Proteomics Standards Initiative	http://www.psidev.info/
MIARE	RNA interference	MIARE Informatics Work Group	http://www.miare.org/
MIBBI[e]	Multi-domain	MIBBI collaboration	http://mibbi.sourceforge.net/
MIFlowCyt	Flow cytometry	Flow Cytometry Consortium	http://flowcyt.sourceforge.net/
MIGS	Genomic sequencing	Genomic Standards Consortium	http://gensc.sf.net/
MIMIx	Molecular interactions	HUPO Proteomics Standards Initiative	http://www.psidev.info/
MIRIAM	Biological modeling	BioModels.net	http://biomodels.net/index.php?s=MIRIAM
MISFISHIE	Gene expression localization	MGED MISFISHIE Working Group	http://mged.sourceforge.net/misfishie/
Proteomics guidelines[c,d]	Proteomics	*Ad hoc* collaboration	http://dx.doi.org/10.1002/pmic.200500856

TABLE 1 (***Continued***)

Product name	Scope	Developers	Link
REMARK	Tumor markers	NCI – EORTC	http://www.cancerdiagnosis.nci.nih.gov/assessment/progress/clinical.html
(ii) Controlled vocabularies and ontologies			
Cell ontology[f]	Cell types	*Ad hoc* collaboration	http://obo.sourceforge.net/cgi-bin/detail.cgi?cell
EXPO	Experiment design	EXPO working group	http://expo.sourceforge.net/
GO[f]	Genes & gene products	Gene ontology consortium	http://www.geneontology.org/
MeSH[f]	Human disease	US National Library of Medicine	http://www.nlm.nih.gov/mesh/meshhome.html
MGED ontology[f]	Transcriptomics	MGED Ontology Working Group	http://mged.sourceforge.net/ontologies/
MSI NMR[f,g]	NMR spectroscopy	Metabolomics Standards Initiative	http://obo.sourceforge.net/cgi-bin/detail.cgi?nmr
OBI[e,f]	Multi-domain	OBI Consortium	http://obi.sourceforge.net/
PATO[f]	Phenotypic features	OBO phenotype group	http://www.bioontology.org/wiki/index.php/PATO:Main_Page
PSI-MI CV[f]	Molecular interactions	HUPO Proteomics Standards Initiative	http://obo.sourceforge.net/cgi-bin/detail.cgi?psi-mi
PSI-Mod[f]	Protein modifications	HUPO Proteomics Standards Initiative	http://www.psidev.info/index.php?q=node/92
PSI-MS CV[g]	Mass spectrometry	HUPO Proteomics Standards Initiative	http://psidev.sourceforge.net/ms/xml/mzdata/psi-ms-cv-latest.obo
sepCV[f,g]	Lab separations	HUPO Proteomics Standards Initiative and Metabolomics Standards Initiative	http://obo.sourceforge.net/cgi-bin/detail.cgi?sep
(iii) Data formats			
AnalysisXML	Mass spec informatics	HUPO Proteomics Standards Initiative	http://www.psidev.info/index.php?q=node/85#analysisXML
AnIML	Analytical instrument data	ASTM International	http://www.astm.org/
CCPN data model	NMR spectroscopy	Collaborative computing project for NMR	http://www.ccpn.ac.uk/
DataXML[h]	Mass spectrometry	HUPO proteomics standards initiative	http://www.psidev.info/index.php?q=node/257
FuGE[e]	Multi-domain	*Ad hoc* collaboration	http://fuge.sourceforge.net/
GelML and GelInfoML	Gel electrophoresis	HUPO proteomics standards initiative	http://www.psidev.info/index.php?q=node/83#formats
JCAMP-DX	Analytical instrument data	JCAMP/IUPAC	http://www.jcamp.org/
MAGE	Transcriptomics	Microarray gene expression data society	http://www.mged.org/Workgroups/MAGE/mage.html
MAGE-Tab	Transcriptomics	Microarray gene expression data society	http://www.mged.org/Workgroups/MAGE/mage.html
MIF (*version 2.5*)	Molecular interactions	HUPO proteomics standards initiative	http://www.psidev.info/index.php?q=node/60
mzData[h]	Mass spectrometry	HUPO proteomics standards initiative	http://www.psidev.info/index.php?q=node/80#mzdata
SBML	Systems biology	The SBML team	http://sbml.org/
spML	Sample processing	HUPO proteomics standards initiative	http://www.psidev.info/index.php?q=node/90#psiFormats

[a] CIMR and MIAPE are actually composed of a series of guidelines modules.

[b] The CONSORT home page links to several other sets of guidelines related to clinical trials.

[c] These guidelines go beyond reporting to actually specify ‘acceptable practices’.

[d] Where no website is available for a project, a relevant publication is given in the rightmost column.

[e] These integrative projects offer (specifically for omics work) a cross-domain solution.

[f] OBO-registered ontologies (http://obo.sourceforge.net/); OBO contains many more ontologies, but not all are mature, or have broad support.

[g] These CVs will ultimately be integrated into OBI.

[h] DataXML will soon replace mzData.

tions consisting of little more than the data themselves, with equal facility (i.e. MI specifications should not be embedded in formats). Most modern data-exchange formats are built, as the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) recommends, using the eXtensible Markup Language (XML), which is really just a way of annotating particular parts of a text-based data file in a computer-readable manner.

There is, of course, a need for software that guides scientists through the process of meeting the requirements of one or more MI specifications, automating data capture where possible (e.g. by importing information directly from instrument control software), enabling straightforward use of CV terms as far as is practical and, perhaps, exporting the collated data set in a standard format (perhaps directly to an appropriate public or private repository similar to [4,9,11–13]). Such software should ensure that scientists are not exposed to any of the various informatics components (as above) lurking 'under the hood'.

To summarize, minimum information requirements should be used to regularize the content of descriptions of protocols and data, controlled vocabulary terms (ideally drawn from well-engineered 'ontologies' to support more powerful downstream querying [14]) should be used extensively in such descriptions, and standards-compliant data formats should be used for the transmission of those descriptions between tools and to repositories, from where they can be accessed at a later date (ideally by leveraging prior use of CV terms). Some of the potential benefits of the standards-based approach have been described already, and more will be offered below, but first, let us consider the potential costs of standards use.

Using the new standards: The costs

The cost of introducing a standards-based approach to data management will be significant, requiring that software and databases be developed or purchased that support use of those standards (formats, CVs and MI specifications alike), that staff are appropriately trained on software, and that person-hours are allocated for what could be termed 'double reporting' (i.e. once in archival mode at regulatory strength using native formats and again to meet the lesser MI specifications in the standard formats using standard vocabularies). It is also the case that were these standards to change significantly over the short-to-medium term, costs would multiply. However, in the opinion of the author, we can rely on the responsible attitude of the developers of these standards, who will normally have committed themselves to regular release cycles and who have, in many cases, established robust relationships with appropriate vendors to encourage the implementation of support (in software) for standards use.

Using the new standards: the benefits

Use of the emerging data formats, controlled vocabularies and especially MI specifications brings a wide range of (mostly efficiency-related) benefits, including

1. *Methods robustly associated with results, in a condensed and accessible form*
   - Data generated by specific techniques or using particular materials can be more straightforwardly identified in repository searches (or *excluded* from results sets)
   - The relevance of data can be assessed through summaries without wasting time wading through full data sets (in various proprietary formats)
   - Data are more likely to survive the obsolescence of old technology
2. *Reduced need to repeatedly construct contextualizing summaries*
   - Facilitates the sharing of data with collaborators
   - Avoids the risk of loss of information through staff turnover
   - Enables time-efficient handover of projects from one researcher to another
3. *Independence from original context (i.e. formats and terminologies from particular software)*
   - Building aggregate data sets containing (similar) data from different sources becomes straightforward; for example, to assess a (generic) protocol through meta-analysis
   - Integrating data from different domains becomes simpler; for example, correlating changes in mRNA abundances, protein turnover and metabolic fluxes in response to a stimulus – *genericized* high-level metadata are absolutely key here
   - Leveraging public data as 'Commercial Intelligence' (CI) becomes useful; that is deriving value from standards-compliant data sets obtained from public repositories (industry should strongly encourage public funders to require that publicly available data are standards-compliant for this very reason, i.e. 'knowledge transfer')
4. *Improving users' experience, reducing burdens on developers*
   - Databases and software that can handle data from various sources become far more straightforward to develop (i.e. data handling and integration are greatly simplified)
   - Software can be developed that leverages the databases described in the previous bullet to help users retrieve data sets from various sources without undue effort.

A choice of standards?

As is evident from Table 1, there are many individual projects in each of the three broad areas described above (MI specifications, data formats and controlled vocabularies). In each case these projects seek to serve particular communities. However, the interests of those communities may well overlap, leading to the situation where there is a choice for third parties as to which 'standard' they use. The situation is most acute for MI specifications: Many of those listed in the first part of Table 1 offer guidance on describing biological material; for example, Core Information for Metabolomics Reporting (CIMR), MIAME and Minimum Information about a Genome Sequence (MIGS). Others overlap on particular technologies; for example, CIMR and Minimum Information about a Proteomics Experiment (MIAPE) both address mass spectrometry. These overlaps need to be resolved through coordination, to the satisfaction of all concerned parties; otherwise we will be faced with a choice of reporting specifications, defeating the purpose of the exercise.

There is a second problem that arises from uncoordinated development of standards: Even if there were no overlaps in scope between like projects, interoperability would still be hindered. Different levels of granularity in description, different approaches to structuring information, or simply the different naming of concepts can confound attempts to combine or search across heterogeneous collections of data sets. (Is the combination of a chromatography column and mass spectrometer a 'platform' or just a set of instruments? Are you running an assay or an experiment or a study? Do you use the language of a physician or a phytologist?)

Left unresolved, these problems would limit the usefulness of these standards in supporting the integration of data sets from different domains. There is a further serious ramification; the tool developers who could provide the kinds of standards-compliant software described above, whether in the commercial or public sector, will not commit resources if the markets for their products are not clearly defined and relatively stable. Communication between projects in support of coordinated development and sharing of best practice is therefore crucial [15]. An early indication that the wider community of standards developers favors such coordination comes from a recent special issue of the journal *Omics*, which was completely given over to the development of standards; it is summarized in [16].

It is clear then that standards development projects in the life sciences need agreed, common frameworks to promote coordinated development and to assist in the identification of overlaps. For example, consider Figure 1, which is a graphical rendering of some of the work of MGED's Reporting Structures for Biological Information (RSBI [17]) working group concerning the basic structure of an investigative project. This structure is application-agnostic (i.e. it is not designed in the style of a format, ontology or MI specification) and has broad applicability in that just about any science discipline, using any kind of analytical technique, can in principle fit within this framework.

Three running projects offer the prospect of development of coordinated standards; one concerned with data formats (FuGE), one with ontology (OBI) and one with MI specifications (MIBBI). Each of these projects is built on collaboration between a broad set of domains and each has open channels by which comments can be passed back to the core developers.

Project #1: FuGE

The Functional Genomics Experiment (FuGE – http://fuge.sourceforge.net/) Project's UML model and XML format constitute a key resource for tool and format developers working in the biological and biomedical sciences. FuGE fulfills two roles: firstly, that of a general model of experimental science capable of describing a vast array of workflows, employing ontologies in a sophisticated manner and 'wrapping up' other data formats and secondly, of a template from which more application-specific formats can be developed, by 'extension' of FuGE classes and relationships. While not heavily supported with tools thus far (except the CPAS software [18], although that uses a pre-release version), FuGE is repeatedly proving its worth in an array of format development projects (e.g. in recent work by the Proteomics Standards Initiative; *cf*. all the PSI formats listed in Table 1 except *mzData*). It is the opinion of the author that any new format development project in the life sciences should look long and hard at FuGE before progressing; it is an excellent piece of design, facilitates more straightforward integration of other data encoded in FuGE-derived formats and enables more efficient tool and repository development through the reuse of code libraries.

Project #2: OBI

The Ontology for Biomedical Investigations (OBI — http://obi.sourceforge.net/) is an integrated ontology that will (once it matures) cater to the needs of a very broad range of life scientists and clinicians in describing the processes of experimental science and clinical investigation. It contains 'universal' terms with broad applicability and domain-specific terms relevant only to particular disciplines (such as environmental biology, transcriptomics or radiology). Note that OBI only covers project structure, methods and to some extent results; the description of organisms, for example, should be done using normal taxonomical resources and ontologies such as PATO (*cf*. Table 1). Although its development is still at an early stage, this ontology shows much promise, largely because of its stated purpose of integrating term sets from many domains; this is key to the future integration of data sets from those various domains in support, for example, of systems biology. Representatives of a large number of communities have already signed up to the project (listed at http://obi.sourceforge.net/community/) and more keep joining all the time, making OBI a strong candidate to become the perfect complement to the FuGE model described above (although, as stated, it is not yet fully mature).

Project #3: MIBBI

MI checklists are usually developed by independent groups representative of particular domains. Consequently, the full range of checklists can be difficult to establish without intensive searching, and tracking their evolution is non-trivial; they are also inevitably partially redundant one against another (e.g. the overlap between MIACA, MIARE and MIFlowCyt described above and most prevalent in relation to the description of biological source material). MIBBI (http://mibbi.sourceforge.net/) simultaneously acts as a 'one-stop shop' for those exploring the range of extant projects and provides a framework within which projects can work together toward gradual integration of their products (ultimately doing for MI specifications what FuGE and OBI will do for formats and ontology).

Conclusions

This review has argued for (1) high-quality annotation of data sets according to agreed specifications, (2) the extensive use of controlled vocabulary terms and (3) the use of standard data formats. The motivation for advocating these principles is to increase the total value derived from scientific investigations. Presently, there are many projects developing resources to support such activities. A significant number of these projects are currently only loosely connected. However, the author confidently expects that over the next two years these many projects will coalesce into groups centered on FuGE, OBI and MIBBI. Movement in this direction is already occurring; the FuGE model continues to grow in popularity, as shown by the increasing number of projects working with the model – several standards bodies now work almost exclusively with FuGE; OBI continues to attract additional communities and has firmly established itself in the ontology community by becoming a candidate for the emerging OBO Foundry (http://obofoundry.org/); and the MIBBI project, behind which a major journal recently threw its weight (author, pers. commun.), is itself growing rapidly and has begun the process of analyzing existing domain-specific MI specifications in advance of generating integrated, orthogonal

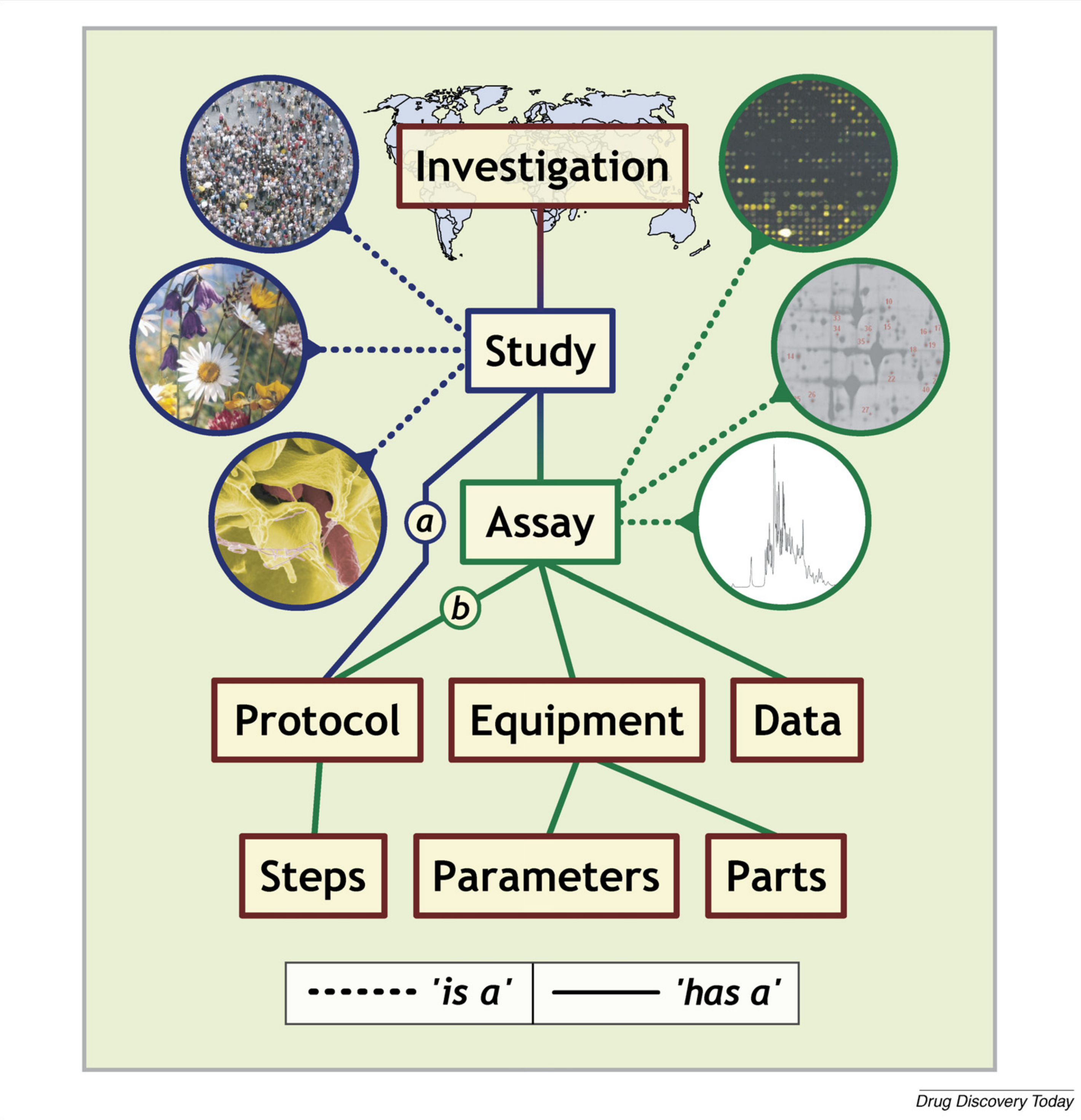

FIGURE 1

A generalized view of the structure of investigative projects in the life sciences. This figure asserts that an '*Investigation*' (of a particular medical syndrome, environmental effect, biological function, etc.) consists of one or more linked '*Studies*' (each of which occurs in the context of a particular biological domain such as toxicology, cell biology or environmental science) that themselves consist of one or more '*Assay*' steps (analysis of part of the material generated or collected for the study, perhaps by use of one or more omics techniques such as proteomics). An '*Assay*' has a '*Protocol*' (through the link marked '*b*') that is itself composed of a series of '*Steps*'; an '*Assay*' involves some '*Equipment*' (which may have '*Parts*' and '*Parameters*'); an '*Assay*' should generate some '*Data*'. This schema could quite easily represent any one of a number of analytical workflows. (N.B. A '*Study*' may also have a '*Protocol*', shown here as the link marked '*a*'; such a '*Protocol*' would relate only to the generation of the material used for the '*Assays*' and would not normally generate data.)

specifications of its own. Furthermore, participants in these three 'meta-projects' regularly interact at meetings and by other means to ensure that each project moves forward in a manner supportive of the other two.

At the point where these three projects can be said to be properly mature, it is inevitable that market demand, encouraged by the growing enthusiasm of both journals [19] and funders (e.g. http://www.bbsrc.ac.uk/support/guidelines/datasharing/context.html) for the standards-based approach to reporting, will drive the production of appropriate free and commercial tools (especially so-called 'electronic note books' [20]) and repositories, simplifying the process of standards-compliant reporting for experimentalists and increasing acceptance and uptake (and therefore, demand) as a result – a virtuous circle. The PSI's *mzData* standard (http://www.psidev.info/index.php?q=node/95) offers an example of this principle at work.

For public science, the widespread use of the standards-based approach offers the prospect of the availability of large quantities of well-annotated, integrable data from an array of disciplines. For the pharmaceutical industry those same public resources will also, of course, be of benefit, but in addition, there is the opportunity to reap the *direct* benefits of the standards-based approach in collecting, managing, integrating and exploiting private data, enabling more efficient development of novel therapeutic agents.

Acknowledgements

Thanks to Dawn Field, Susanna-Assunta Sansone and Norman Morrison for useful discussions, and to the referees and editors for constructive feedback.

References

1 Anon., (2006) Under the MIAME sun. *Nat. Methods* 3, 415

2 Brooksbank, C. and Quackenbush, J. (2006) Data standards: A call to action. *Omics* 10, 94–99

3 Burgoon, L. (2006) The need for standards, not guidelines, in biological data reporting and sharing. *Nat. Biotechnol.* 24, 1369–1373

4 Edgar, R. and Barrett, T. (2006) NCBI GEO standards and services for microarray data. *Nat. Biotechnol.* 24, 1471–1472

5 Kenter, M.J. and Cohen, A.F. (2006) Establishing risk of human experimentation with drugs: lessons from TGN1412. *Lancet* 368, 1387–1391

6 Hogan, J.M. *et al.* (2006) Experimental standards for high-throughput proteomics. *Omics* 10, 152–157

7 Plint, A.C. *et al.* (2006) Does the CONSORT checklist improve the quality of reports of randomized controlled trials? A systematic review. *Med. J. Austr.* 185, 263–267

8 Smidt, N. *et al.* (2006) The quality of diagnostic accuracy studies since the STARD statement – Has it improved? *Neurology* 67, 792–797

9 Brazma, A. and Parkinson, H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat. Biotechnol.* 24, 1321–1322

10 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371

11 Jones, P. *et al.* (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 34 (Database issue), D659–D663

12 Hermjakob, H. and Apweiler, R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev. Proteom.* 3, 1–3

13 Spasić, I. *et al.* (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinform.* 7, 281

14 Gardner, S.P. (2005) Ontologies in drug discovery. *Drug Discov. Today: Technol.* 2, 235–240

15 Ball, C.A. (2006) Are we stuck in the standards? *Nat. Biotechnol.* 24, 1374–1376

16 Field, D. and Sansone, S.-A. (2006) A special issue on data standards. *Omics* 10, 84–93

17 Sansone, S.-A. *et al.* (2006) A Strategy Capitalizing on Synergies: The Reporting Structure for Biological Investigation (RSBI) Working Group. *Omics* 10, 164–171

18 Rauch, A. *et al.* (2006) Computational proteomics analysis system (CPAS): An extensible: open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* 5, 112–121

19 Anon, (2006) Standard operating procedures. *Nat. Biotechnol.* 24, 1299

20 Taylor, K.T. (2006) The status of electronic laboratory notebooks for chemistry and biology. *Curr. Opin. Drug Discov. Dev.* 9, 348–353